Environmental Health Perspectives
Vol. 32, pp. 267-271, 1979

Statistical Approaches to Toxicological Data

by D. G. Hoel*

Statistical techniques as applied to toxicological data are discussed. Issues concerning statistical hypothesis testing and combining studies are considered as well as design of experiments. The problems surrounding risk assessment are also mentioned.

Introduction

In enumerating the objectives of statistics in toxicology, probably the most commonly recognized one is the attempt to quantify the precision of various estimates. Typically, this is concerned with quantifying one's confidence in an observed treatment control difference, and possibly with the confidence one has in an observed dose-response relationship.

From this point of estimation of precision, one moves logically into the area of combining conclusions from various studies in order to numerically describe the effects of a given agent or agents. In this instance we must consider doing more than simply tabulating the positive and negative studies. The statistician can be of assistance by examining the data and explaining inconsistencies if they do, indeed, exist. Very often when one considers the differences of dose and duration of studies and the statistical power of different assays, there may be more consistency than is apparent by simply categorizing the studies with regard to whether or not they achieve an arbitrary 5% level of statistical significance.

Another area of statistics in toxicology which receives a modicum of attention, surely not as much as it merits, is the design of experiments. We are referring to the classical fundamental question of dose selection and sample size. In working with toxicologists, one often finds that studies are poorly or insufficiently designed. Some of the time this is excused on the basis of doing only a "pilot study." Often the projected study is never conducted and the casually designed pilot work stands as is, in its incomplete state. Thus, I feel toxicologists should more often consult with the statistician initially. In this way they can clarify and define their objectives, and hopefully, this early interaction of statistician and toxicologist will lead to more efficient and sensibly designed studies.

The topic of data reduction and interpretation is also of more than passing interest. Specifically, we refer to reducing rather complicated, multifactor, multiresponse data into more easily understandable terms. One example of this is experimental information related to multivariate time series data sets which is being generated by some types of behavioral studies. I believe that this area has an important future in biostatistics and toxicology.

The next aspect of statistics which has been receiving much attention in toxicology is the construction and application of mathematical models in order to better understand biological processes. Considerable effort has been directed not so much toward understanding the processes, but in the direction of trying to extrapolate processes outside their range of observation. Recently, this approach has been utilized in an attempt to understand the potential risks of low dose exposure levels to man, based upon the extrapolation of high dose toxicological data. In this area, statistical errors are being discussed, but we have not as yet been able to satisfactorily quantify the biological errors. In other words, what are *implications of faulty model assumptions*?

The final area of concern to statisticians is the application of operations research techniques. From regulations such as the Toxic Substance Act we are

*Biometry Branch, NIEHS, Research Triangle Park, North Carolina 27709

delving further into questions of how to best manage batteries of tests. For example, we must consider which short term tests should be used, and in what combinations, in order to yield a total toxicological evaluation which is resource efficient.

Statistical Hypothesis Testing

The first area I would like to address is the problem of statistical hypothesis testing in bioassay work. This is probably the statistical issue which presents itself most frequently and is typically concerned with whether or not the 5% level was achieved in terms of a treatment control comparison in tumor incidences, for example. Here we are concerned with the type of error level associated with a false positive. Often if a 5% level is achieved, we feel that we have a significant result. We further arbitrarily assume that 5% to 10% is suggestive, and greater than 10% is nonsignificant. We tend to classify this way, and these differences in carcinogenesis clearly depend upon dosage levels, the number and size of the experimental groups, the duration of the study and so on. I feel that there is a greater need to assess and understand the power of the assays. If you look at experimental carcinogenesis, for example, with a standard 100 treatment, 50 control animals, and consider a spontaneous 5% level of tumor incidence and compare that with an increase of about 5%, you will find that you do not have very good power of detection (less than 5% chance) (*1*). Now, with this power information, the investigator may decide not to proceed with his design. On the other hand he may want to go ahead because he can obtain some sort of upper bound on the degree of a possible increased incidence rate. This leads to the problem of combining tests. The problem of statistically classifying tests by their significance; that is, whether each is positive, negative, or suggestive, leads to apparent inconsistencies when, in fact, there may biologically be none. How can we better approach this? One method, of course, is to attempt an estimation of the effect; a quantification of the effect with some measure of confidence. For example, epidemiologists typically talk about relative risk and the confidence associated with the risk estimate. In comparing several studies, they are able to compare the relative risks and, in fact, numerically combine them. Statisticians can also become involved with models as they try to compare studies. Meselson (*2*) did an interesting study comparing particular compounds which had carcinogenic effects in animals and also exhibited some evidence of human cancer effects. He was especially interested in comparing a measurable potency or degree of carcinogenicity between rodent species and man. Now, in making such comparisons by looking at various studies, there are models that one can use to predict the effects of lifetime carcinogenesis exposure based upon less than lifetime exposures. For example, many of the rodent studies were lifetime studies, while the human data were typically less-than-lifetime occupational data. It has been suggested by Armitage and Doll (*3*) that the cancer incidence rate is proportional to the duration of exposure raised to a power equal to one less than the number of states which represent the initiation process.

Techniques of this sort can be used to compare and bring varying studies together to hopefully form a more complete picture of the toxicological effect. It has been proposed recently that with carcinogenicity studies we report a potency value such as a cancer dose 50. This approach has been considered by several investigators (*4, 5*) and is quite useful when comparisons are attempted between differing assays such as the Ames Salmonella assay and chronic carcinogenesis studies. With the Ames test we are concerned with the dose, which yields twice the number of revertents than would have been observed spontaneously. One then compares that dose with some measure of carcinogenicity potency from the rodent studies. So, we could discuss what dose would induce tumors in 50% of a particular rodent strain, and this would in turn give us some measure of cancer potency to compare to the mutagenicity dose. Of course, a single number does not adequately describe the entire shape of the dose response function, so care must be taken. In any case, I strongly feel we need to work towards integrating studies. Closely related to this issue is the problem of multiple tests in a single study. For example, if one were conducting typical teratological assays with examinations for various types of anomalies, an increase in cleft palates, say, may be observed. This statistical increase may be the result of much simultaneous testing of various types of anomalies. Salsburg (*6*) referred to this in terms of the cancer bioassays. His concern is the situation which exists when you are looking at a number of different tumor types. If you have independent tests, which often times you do not, the error rate of false positives may be much higher than expected. The NCI bioassay historical information on spontaneous rates suggests that the actual individual error rates are quite small due to the low spontaneous incidence rates. Thus, with low actual error levels, the overall false positive rate is reasonable. This has been discussed by Fears, Tarone, and Chu (*7*) with emphasis on particular strategies for combining comparisons from various doses, species, and sexes. Using reasonable strategies they show that you were probably dealing with the error level that you thought you were, or at least approximately so, in spite of the many multiple comparisons. This is interesting when considered in light of Bayesian procedures. By dependency upon certain historical information in terms of spontaneous rate for the strains included in the assay, we are using *a priori* information on the probability of a spontaneous outcome in order to control the nominal error rates. Clearly, much work is needed to understand what error levels are involved and how best to utilize available historical information.

Besides carcinogenesis there are many other special methodologies for specific bioassay in toxicology. For instance, how one should analyze teratological data or dominant lethal assays when dealing with litters is not clear. It does seem apparent, however, that one should not analyze on a per fetus basis since the dam is the sampling unit. There are a number of models and procedures which have been developed and deal with such diverse methods as u-tests on proportions of affected fetuses with censoring and possible jackknifing, and setting up models such as a beta-binomial representation of inter-litter dependency. In sum, it is not yet clear which, if any, method is best for treating the data. In fact, this is an area where we need to focus on comparing these procedures and seeing if it makes any practical difference whether or not we are using an optimal statistical procedure. So far it seems that the only general conclusion which has come through clearly is to avoid using a chi-square on numbers of affected fetuses.

For lifetime carcinogenesis studies, it is usually assumed in analyzing the data that we are employing competing risk and life table techniques. These methods require an assumption of independent risks of death. Neyman and colleagues at Berkeley (*8-10*) have recently taken a scrutinizing look at this problem and have shown that this may not be a reasonable approach and have proposed using Markov models instead. In this instance we refer to disease states and Markov transitions between the disease states. The issue is, if we use competing risk techniques, how robust are they if the assumption of independent diseases is incorrect? The alternative to the application of the Markov disease state model to data also presents problems. Are there sufficient parameters for the model to be realistic? If so, can we analyze the data from very large studies with serial sacrifice with many Markov parameters? Also, in dealing with carcinogenesis data, we have problems with specifying a cause of death. As was mentioned earlier, the analysis requires that cancer in animals is classified as either the cause of death or incidental. This toxicological classification in addition to mixing in serial sacrifice data with the normal mortality, causes considerable difficulties in developing efficient methods of analysis. Quite clearly we still have serious problems in the analysis aspect of carcinogenesis studies. How great our problems are, we do not know. It may be that our current unrefined techniques are sufficient, but we cannot feel assured of this until further studies have been completed.

Experimental Design

Now, moving from the analysis of bioassay data to the design of experiments in the carcinogenesis area, we have been mainly looking at high dose screening studies to determine whether or not a given compound is carcinogenic. We also have need to understand the dose-response distribution if we want information concerning low dose risk estimation. From a design standpoint we do not know how to combine dosages to obtain a balance between the two needs. As we add low doses, we of course give up some of the ability of the assay to detect the effects. In other words, we must often take away the sample size from the high dose levels since the lifetime study is such that we do not have the time to avoid making decisions until more low dose information is available. It may be possible on the basis of short term assays to get an indication of the likelihood of the material to indeed be carcinogenic, some idea about how potent it might be and, thus, where we should locate the design points. But, to design well, we must have some prior knowledge of potency.

Some design work has been done for low dose extrapolation. If one is interested solely in adding one experimental dose and drawing a straight line, assuming no background and working with confidence intervals, Crump and Langley (*11*) have shown that approximately the best place to locate the experimental dose would correspond to that dose which is one-half the dose which would yield a 10% incidence. Recently there have been more elaborate attempts to make use of models such as the multistage and to design optimum dose levels in terms of sample size and dose selection, depending on the range of parameters in the model and on the number of stages. In all of this design work we require some prior knowledge concerning potency, and we have to work towards a balance between screening versus the knowledge of the dose response function.

Sequential design techniques have not been utilized often in toxicology as compared with, for example, clinical trials. A few examples do exist, one being the work of Generoso (*12*), in screening for heritable translocations. Sequential breeding schemes have been developed. These detect semisterile mice with certain statistical confidence based on breeding performance, instead of resorting directly to expensive cytology. A related design problem is concerned with research questions of which tester strains one should select from the Ames assay, and what other short term test should be employed for screening, if the end point is carcinogenesis. In other words, we are asking what battery of tests and in what sequence a toxicological screen should be composed. Currently, several government agencies are dealing with this issue. Hopefully, the solution will depend upon both biological theories and considerable empirical evidence. For example, compounds which are carcinogenic and their noncarcinogenic chemical analogues are being tested in laboratories throughout the world using various short term tests. Finally, there is the possibility of making the individual analyses more efficient. Recently, Fears and Douglas (*13*) considered schemes for more efficient use of pathologists' time in carcinogenesis studies. For example, one can envision a scheme whereby pathologists initially look at only the high dose groups and work with gross pathology before proceeding to more expensive microscopic work. All of these examples suggest design approaches which can provide substantial savings of limited toxicological testing resources.

Risk Estimation

Recently, interest has been directed to statistical models of dose response for estimating low dose risk based upon high dose data. Curve fitting has been commonplace using functions such as probit, logit, multihit, multistage, etc. Some of these functions have a tradition in bioassay work, while others attempt crudely to describe biological mechanisms. For example, the multistage model assumes the carcinogenesis process is represented by a direct-acting carcinogen interacting with DNA as in a single cell somatic mutation theory. This simple model does not include consideration of DNA repair, the immune surveillance system, genetic or environmental susceptibilities in the population, or pharmacokinetics.

Currently, pharmacokinetic models are being incorporated into the risk estimation process. For example, Gehring (*14*) has been conducting some interesting work with vinyl chloride. He has determined the Michaelis-Menten parameters and applied these to exposure levels in Maltoni's vinyl chloride carcinogenesis studies. This dose adjustment more or less straightens out the dose response so that effect is proportional to dose. Theoretical models are quite useful, because one can postulate various mechanisms and then examine how effective they are in relation to issues such as risk assessment. For example, in pharmacokinetics we want to know how critical it is when one is estimating low dose effects. Can the incorporation of kinetics change estimates by orders of magnitude or not? It certainly is real mechanistically, but quantitatively we must determine how important pharmacokinetic considerations are when decisions on allocation of toxicological resources are made. These are very difficult resource questions that have to be worked out jointly with toxicologists, biologists, and mathematicians.

From theoretical considerations, we can say that we do know that one cannot distinguish between the models if the response is purely dichotomous (i.e., cancer or no cancer). Also, we cannot establish from the data alone either the existence or nonexistence of thresholds, nor can we rule out the presence of a linear term which predominates at the low dose levels. Also it is well known that if risk estimates are made very far from the experimental region, then the choice of models is quite critical.

Potentially the greatest errors in extrapolating are associated with mouse-to-man extrapolations. Much empirical work for assessing the variability between strains and species from existing data sets is needed. Hopefully such empirical studies will help to provide information about the statistical confidence in quantifying the biological errors associated with risk estimation.

In conclusion, I see in various toxicological problem areas attempts to build mechanistic models and then to evaluate the available statistical procedures. Also, there is a need to expend a greater effort in trying to evaluate the empirical evidence for consistency with models and for attempts to quantify biological model errors. Also, we must begin attempts to develop efficient strategies for toxicological assessment of materials. This applies to both the different assays we have available and to using related information such as chemistries, food consumption, etc., associated with these assays, which is typically ignored statistically. Finally, in all aspects of statistical design and analysis of toxicological experiments, we need a greater emphasis on the understanding of the biological mechanisms involved. With proper collaboration, both the toxicologist and the statistician will benefit.

REFERENCES

1. Hoel, D. G. Some problems in low-dose extrapolation. In: Origins of Human Cancer, Book C. H. H. Hiatt, J. D. Watson, and J. A. Winsten, Eds. Cold Spring Harbor Laboratory, Cold Spring Harbor, 1977.
2. NAS, Contemporary pest control practices and prospects. In: Pest Control: An Assessment of Present and Alternative Technologies. Volume I. National Academy of Sciences, Washington, D.C., 1975.

3. Armitage, P., and Doll, R. Stochastic models for carcinogenesis. In: Proceedings of the Fourth Berkeley Symposium, University of California Press, Berkeley, Calif.
4. McCann, J., and Ames, B. N. The Salmonella/microsome mutagenicity test: predictive value for animal carcinogenicity. In: Origins of Human Cancer, Book C. H. H. Hiatt, J. D. Watson, and J. A. Winsten, Eds. Cold Spring Harbor Laboratory, *Cold Spring Harbor, 1977.*
5. Meselson, M., and Russell, K. Comparisons of carcinogens and mutagenic potency. In: Origins of Human Cancer, Book C. H. H. Hiatt, J. D. Watson, and J. A. Winsten, Eds. Cold Spring Harbor Laboratory, Cold Spring Harbor, 1977.
6. Salsburg, D. S. Use of statistics when examining lifetime studies in rodents to detect carcinogenicity. J. Toxicol. Environ. Health 3: 611 (1977).
7. Fears, T. R., Tarone, R. E., and Chu, K. C. False-positive and false-negative rates for carcinogenicity screens. Cancer Res. 37: 1941 (1977).
8. Neyman, J. Assessing the chain: energy crisis, pollution and health. Int. Stat. Rev. 43: 253 (1975).
9. Clifford, P. Nonidentifiability in stochastic models of illness and death. Proc. Natl. Acad. Sci. 74: 1338 (1977).
10. Tsiatis, A. A nonidentifiability aspect of the problem of competing risks. Proc. Natl. Acad. Sci. (U.S.) 72: 20 (1975).
11. Crump and Langley
12. Generoso, W. M., *Russell*, W. L., *and Gosslee*, D. G. A sequential procedure for the detection of translocation heterozygotes in male mice. Mutat. Res. 21: 220 (1973).
13. Fears, T. R., and Douglas, J. F. Suggested procedures for reducing the pathology workload in a carcinogen bioassay program, Part I. J. Environ. Pathol. Toxicol. 1: 125 (1977).
14. Gehring, P. J., Watanabe, P. G., and Park, C. N. Resolution of dose-response toxicity data for chemicals requiring metabolic activation: example — vinyl chloride. Toxicol. Appl. Pharmacol. 44: 581 (1978).